Operational comparison of rainfall-runoff models through hypothesis testing

James Fidal[1]; and Thomas Kjeldsen[2],

ABSTRACT

Assessing rainfall-runoff model performance and selecting the model best suited are important considerations in operational hydrology. However, often model choice is heuristic and based on a simplistic comparison of a single performance criterion without considering the statistical significance of differences in performance. This is potentially problematic as interpretation of a single performance criteria is subjective to the user. This paper removes the subjectivity by applying a jackknife split-sample calibration method to create a sample mean of performance for competing models which are used in a paired t-test allowing statements of statistical significance to be made. A second method is presented based on a hypothesis test in the binomial distribution, considering model performance across a group of catchments.

A case study comparing the performance of two rainfall-runoff models across 27 urban catchments within the Thames basin show that while the urban signal is difficult to detect on single catchment, it is significant across the group of catchments depending upon the choice of performance criteria. These results demonstrate the operational applicability of the new tools and the benefits of considering model performance in a probabilistic framework.

Keywords: Comparison techniques, Hypothesis test, Jackknife split-sample, Hydrological model, Uncertainty analyse, Operational, Statistical significance.

INTRODUCTION

Hydrological modelling plays a key role in water management with many practical appli-

[1]Dept. of Architecture and Civil Engrg., Univ. of Bath, England, BA2 7AY. E-mail: jf643@bath.ac.uk.
[2]Dept. of Architecture and Civil Engrg., Univ. of Bath, England, BA2 7AY.

cations such as extending stream flow records, predicting future river flows, and simulating river flows in ungauged catchments (Beven 2011). Addor and Melsen (2019) provide an insight into model selection comparing 1,529 abstracts and found that the choice of models can be predicted based on the first author in 74% of studies, hence models are typically selected based on familiarity as opposed to the most adequate model. However the results are from academic studies only, and did not consider industry. Fleming (2009) performed a more in depth analysis by quizzing 47 hydrological professionals from academia (24%), government (47%) and private sector (29%) about model use and selection with the reputation of the model and performance listed as key reasons for selection. The results from both Addor and Melsen (2019) and Fleming (2009) present interesting conflicting conclusions that model selection is based on familiarity in academia and previous model performance and reputation in government and private sector.

Deciding if a particular model can adequately represent observed data is typically based on a performance criteria such as the Nash-Sutcliffe efficiency (NSE), Root Mean Square Error (RMSE) or the coefficient of determination (R^2), derived by comparing observed and simulated runoff. Legates and McCabe (1999) argued that simply applying and presenting these criteria is too simple and potentially misleading basis for model selection. They compared a number of performance criteria and concluded that they can be misleading due to sensitivity to extremal values and insensitive to additive and proportional differences between model simulation and observed data. They concluded that performance criteria should ideally be used in conjunction with other methods to evaluate model performance. Issues arising when comparing model performance using performance criteria were also explored by Schaefli and Gupta (2007) who concluded that simply relying on the Nash-Sutcliffe efficiency alone is not sufficient to validate a model, as difference in performance at low flows and peaks are lumped together and the performance cannot be captured within a singular value. Similarly Krause et al. (2005) argued that no single performance criteria can be considered singularly 'best' as advantages and disadvantages are evident for all criteria. For

example, the Nash-Sutcliffe efficiency and the coefficient of determination are particularly sensitive to model performance at peak flows. Weglarczyk (1998) highlighted the dangers of using multiple performance criteria due to the interdependence between them. To overcome some of these problems alternative performance criteria have been proposed such as the Kling-Gupta efficiency, which is an equal weighting of three components; (i) correlation, (ii) bias, (iii) and variability measures (Gupta et al. 2009).

Similar to these studies Vogel and Sankarasubramanian (2003), Mishra (2009) and Pechlivanidis et al. (2010) all concluded that selecting one single performance criteria, or in some cases using several performance criteria, to determine model performance can be misleading, and that alternative methods such as: hydrograph analysis, covariance validation procedures or uncertainty analysis techniques should be used as well. Uncertainty is a key issue and is prevalent throughout all stages of hydrological modelling, not just model performance, including: input data, model parameter quantification, and model structure as explored by Kavetski et al. (2006), Shen et al. (2012) and Vrugt et al. (2003). Pappenberger and Beven (2006) provided a commentary on the importance of integrating uncertainty analysis into modelling studies, but highlighted that it is not commonly used due to a lack of guidance on the methods to use.

Studies have attempted to incorporate alternate performance comparison methods into model assessment studies, such as Bouffard (2014) who compared two models (TOPMODEL and HBV) using both performance criteria and uncertainty analysis. The performance criteria used was a fuzzy measure combining the Nash-Sutcliffe efficiency, logged Nash-Sutcliffe efficiency and volumetric error. The uncertainty analysis was based on the generalized likelihood uncertainty estimation (GLUE) framework. Anh et al. (2010) compared three models (MIKE-FEH, NAM and TVM) using three different performance criteria (NSE, RMSE and coefficient of determination R^2) and graphical analysis on a single catchment. Both Bouffard (2014) and Anh et al. (2010) provide reasoning behind why each performance criteria was selected, highlighting that different types of performance criteria are sensitive to different

aspects of flow. Refsgaard and Knudsen (1996) used a combination of performance criteria (flow duration curve error index EI and R^2) and plotting observed and model simulated data to compared three models (NAM, MIKE SHE and WATBAL) on three different catchments. Whilst all of these studies have applied various techniques to strengthen the scientific basis for model selection, they all rely on a direct comparison of performance criteria. In summary, model selection based on performance criteria is widespread, but the need for acknowledging uncertainty is evident. Therefore, new methods need to be developed in order to expand current applications of performance criteria to address short-comings yet remain operationally useful.

Kirchner et al. (1996) argued as a minimum, model performance should involve three key elements: (i) a performance criterion,(ii) a benchmark model, against which other models are being tested, and (iii) an assessment of how much better or worse the alternative models perform against the benchmark model. Schaefli and Gupta (2007) also suggested that a simple benchmark model should be applied when evaluating performance criteria. Seibert et al. (2018) proposed upper and lower benchmarks as opposed to a single model to compare against. Whilst methods exist to address point (iii) from Kirchner et al. (1996), this paper will develop two statistical testing frameworks to explore model performance differences.

In order to apply the model selection tools developed within this paper, a suitable calibration method needs to be defined. Klemeš (1986) proposed four methods to calibrate models, the first by splitting the available data into a calibration and validation period, called the split-sample test. This method is now a standard in most hydrological model applications (Andréassian et al. 2009), evidenced from studies such as Refsgaard (1997), Ewen (2011), Donnelly-Makowecki and Moore (1999) and many others. However, as argued by Klemeš (1986), the differential split-sample test, should be required when using a hydrological model to simulate flow within a gauged basin. The differential split-sample is similar to the split-sample test but the calibration data is chosen based on climatic differences. The differential split-sample method is used in hydrology through studies such as: Seibert (2003), Xu (1999),

Donnelly-Makowecki and Moore (1999) and Refsgaard and Knudsen (1996), but it is not as prominent as the split-sample test. The third and fourth method proposed by Klemeš (1986) is to use the split sample and differential split-sample tests on proxy-basins. This would require calibration on catchment A followed by validation on catchment B. However the third and fourth methods are less wide-spread in the literature but have been adopted in studies such as Donnelly-Makowecki and Moore (1999), Santos et al. (2018) and Refsgaard and Knudsen (1996). Reasons for why the third and fourth method are not used were explored by Andréassian et al. (2009) who argued that since the proxy-basin methods generally provide much lower performance scores, modelers are adverse to using these methods. This conclusion is echoed by Seibert (2003).

Research has been done to expand upon calibration strategies such as Ewen and O'Donnell (2012), who explored the idea of improving split-sample calibration and validation by splitting the calibration period into two sections instead of one: a first calibration period, a second calibration period, and finally a validation period. They concluded that whilst this methodology did improve model simulation results, further research was still needed to determine if the methodology will work on different catchments and different types of storms.

Gharari et al. (2013) presented an alternative method applicable to a singular calibration period by splitting the calibration period into a number of sub-periods and generating multiple parameter sets. The performance of each parameter set is then calculated and compared in order to determine the single best parameter set. This methodology only generates multiple calibration periods not validation periods. However, as Gharari et al. (2013) discus the method presented was not a method of addressing parameter uncertainty but was intended to build upon the traditional Klemeš (1986) split sample tests. Coron et al. (2012) expanded upon the split-sample test, called the generalized split-sample test. This test splits the calibration data set into smaller overlapping calibration periods so multiple calibrations are achieved on the same data set. Coron et al. (2012) noted an advantage of the procedure being a large number of results to be analysed. However the problem with this approach is

each validation period is not independent due to sample reuse (Kohavi et al. 1995).

An advancement to calibration-validation methodologies is presented in this paper based on a jackknife style methodology. Building upon (i) the ideas used to determine parameter variability introduced by Jones and Kay (2007) and Selle and Hannah (2010), and (ii) splitting the calibration period into subsets (Gharari et al. 2013) and (Coron et al. 2012) this paper will introduce a new jackknife methodology to calibrate and validate two different models in order to develop a more robust calibration/validation methodology. Whilst a number of studies explore alternative methods of model performance beyond a single performance criteria the use of these criteria is generally accepted in hydrology as a convenient form of comparison with nearly every study using it. This paper presents two new easy-to-use methodologies to measure the difference in performance between hydrological models through a hypothesis-testing framework. These tools will be reliant on existing and widely used performance criteria. The first method is a performance measure formulated as a paired t-test for analysing the performance between two models on individual catchments. The second method is based on a binomial distribution, and tests for statistical significance between model performance across a group of catchments. Both tests rely on non-parametric jackknife resampling to quantify the uncertainty of model performance.

MODEL COMPARISON TECHNIQUES

Model calibration and Validation periods

The first step needed to determine model performance is to define a calibration and validation period. The calibration period is defined as the span of observed data used for calibration of model parameters. The validation period is defined as the period of data in which an independent comparison of observed and simulated data is undertaken to determine if the model is capable of making accurate simulation when applied outside of the calibration period. Calibration and validation periods which do not overlap are used, as a model needs to be validated on data independent of calibration data in order to show that it has the capacity to predict data and not simply mimic calibration data.

Traditionally calibration periods are longer than validation periods as models need to train parameters to estimate the optimal parameters. However one problem with this method is that only a single performance criteria can be obtained. The methods presented here will generate multiple performance criteria from which a mean and standard deviation of the performance criteria can be estimated and subsequently used in hypothesis testing. However this new method requires multiple shorter calibrations than is traditionally used combined with much longer validation periods.

Jackknife calibration method

The jackknife re-sampling technique developed by Quenouille (1956) is a systematic sampling method, which was originally designed for exploring bias estimation but can also provide estimates of variance; the primary reason for adopting the method in this study.

The jackknife method used here is based on the approach proposed by Jones and Kay (2007) to quantify model parameter uncertainty but adapted here as part of the calibration and validation to assess uncertainty of the performance criteria. The jackknife methodology will generate multiple performance criteria for two different rainfall-runoff models applied to a single catchment. A paired t-test method will then be applied to these multiple performance criteria to determine if a significant difference in model performance between the two models can be detected for a given catchment. This process will be repeated for multiple catchments, and a second a binomial hypothesis test method will be applied to test significance between the difference in performance of two models across multiple catchments. The method is presented in the bullet list below:

- Denote M_1 as a model that requires parameter calibration, using a set of observed hydrological data (runoff, rainfall, potential evaporation) of length N.
- Split the data set into $j = 1, ..., n$ equal length (N/n) non-overlapping periods. Each of these periods will be used to calibrate the model, such that the first sub-period has the same first value as the full set and the final sub-period has the same final value

as the full set.

- Calibrate M_1 using the data in the first sub-period ($j = 1$), resulting in a set of model parameters θ_1.
- Use model M_1 with parameter set θ_1 to simulate runoff on the remaining data without the first sub-period, i.e the validation period. A performance criteria Z_1 can be obtained for the validation period by comparing the model simulated data and observed data in the validation period.
- The model is calibrated on the second sub-period, $j = 2$ to obtain a new parameter set θ_2.
- Use M_1 with parameter set θ_2 to simulate runoff on the remaining data without the second sub-period, i.e validation period. A performance criteria Z_2 can be obtained for the validation period.

This process is then repeated systematically until the model have been calibrated on all individual sub-periods, each time using a different sub-period for model calibration and validation. Note that the validation data length is the same each time. For each iteration, a single performance criteria Z_j and parameter set θ_j are obtained, such that a set of $Z_j, j = 1, ..., n$ performance criteria and $\theta_j, j = 1, ..., n$ parameter sets are obtained, with n being the number of non-overlapping periods. Figure 1 shows an example of the jackknife process for a data set of length 30-years, a calibration period of 2-year duration, a validation period of length 28-years; hence a total of 15 performance criteria ($Z_j, j = 1, ..., 15$) would be obtained.

<Figure 1 >

Finally the set of performance criteria $Z_j, j = 1, ..., n$ can be used to assess the uncertainty in model performance. The next two sections will outline how the performance criteria sets obtained can be used to evaluate model performance through standard statistical hypothesis tests.

Paired t-test method

The first method to evaluate model performance is based on the concept of a paired t-test, by comparing two sets of performance criteria obtained by applying two different models M_1 and M_2 to the same data set. The jackknife methodology, described in the previous section, is combined with two models M_1 and M_2 to obtain two sets of performance criteria $Z_{1,j}$ and $Z_{2,j}$ such that both sets are of equal length with the first subscript indicating which model it is obtained from (1 or 2) and subscript j refers to sub-period. The difference in model performance for each sub-period is given as $Z_{d,j} = Z_{1,j} - Z_{2,j}, j = 1, ..., n$. The mean of Z_d can be obtained as:

$$\overline{Z_d} = n^{-1} \sum_{j=1}^{n} Z_{d,j}. \tag{1}$$

The associated jackknife variance estimate $Var\{\overline{Z_d}\}$ is given by Efron (1982), and due to systematically re-sampling from the same data set an inflated variance must be used; i.e.

$$Var\{\overline{Z_d}\} = \frac{n-1}{n} \sum_{i=1}^{n} (Z_{d,j} - \overline{Z_d})^2. \tag{2}$$

According to the central limit theorem the mean value $\overline{Z_d}$ is assumed normal distributed, and consequently the confidence interval of the mean performance measure $\overline{Z_d}$ of each model defined as

$$\left(\overline{Z_d} - z_{(1-\frac{\alpha}{2})} \sqrt{Var\{\overline{Z_d}\}}, \overline{Z_d} + z_{(1-\frac{\alpha}{2})} \sqrt{Var\{\overline{Z_d}\}} \right). \tag{3}$$

The value of $z_{(1-\frac{\alpha}{2})}$ is the $(1 - \frac{\alpha}{2})$ quantile of the standard normal distribution. The 95% confidence is defined for $\alpha = 5\%$ which gives $z_{(0.975)} = 1.96$. A hypothesis test can be formed such that the null hypothesis (H_0) states that the performance of models M_1 and M_2 is the same, i.e

$$H_0 : \overline{Z_d} = 0 \tag{4}$$

The alternative hypothesis can be defined as either model performance is different (two-tailed) or to test one model outperforming the other (one-tailed). An assumption can be made when testing if one model is expected to perform better than another. The case that will be presented in this paper will be the two-tailed test i.e either model is performing better than the other. This is important because at the individual catchment it can not be assumed that either model outperforms the other.

$$H_1 : \overline{Z_d} \neq 0 \tag{5}$$

In order to determine which hypothesis to accept and reject, the confidence interval for $\overline{Z_d}$ is interpreted such that if the interval contains zero then the null hypothesis of equal performance can be accepted, whereas if the interval does not contain zero then the null hypothesis can be rejected, thus indicating a significant difference in model performance.

Binomial hypothesis test

Whilst the previous section outlines a method to compare models on a singular catchment, the binomial hypothesis test is introduced here to compare model performance across a group of catchments. The test will utilise a success/ failure approach to compare two models; either a model outperforms the other or it does not. In order to apply a binomial hypothesis test, independence between model performance assessments in each sub-period has to be assumed, meaning each calibration of a model must be independent and each calibration must not be influenced via previous calibrations.

Consider a region consisting of $i = 1, ..., c$ catchments. Each catchment has a record length N_i which result in a set of sub-periods n_i. Applying the jackknife calibration/validation method to two models M_1 and M_2 on each catchment in turn will result in c sets of performance criteria each containing n_i elements $Z_{1,j,i}$, $j = 1, ...n$, $i = 1, ..., c$ and $Z_{2,j,i}$, $j = 1, ...n$

, $i = 1, ..., c$, with subscript i denoting the catchment number. For each catchment the difference between model performance is calculated for each of the $j = 1, .., n_i$ validation periods, and the difference $Z_{d,j,i}$ can be obtained ($Z_{d,j,i} = Z_{1,j,i} - Z_{2,j,i}$). The means of each of the difference sets can be obtained for each catchment $\overline{Z_{d,i}}$. This process is shown in Figure 2.

<Figure 2 >

The hypothesis test starts from the premise that if the two models perform equally well then there will be a 50-50 chance that one model outperforms the other on any given catchment. The hypothesis test can then be formed such that the null hypothesis (H_0) states that M_1 performs better than (or equal to) M_2 across c catchments, and that the probability this is true is set at less than or equal to a half, (i.e not be chance)

$$H_0 : p \leq 0.5. \tag{6}$$

The alternative hypothesis (H_1) is that M_2 perform better than M_1 across c catchments

$$H_1 : p > 0.5. \tag{7}$$

For convenience, a trial is defined as an extreme based on the mean difference of performance criteria for two models $\overline{Z_d}$ on a particular catchment. Define a successful outcome of a trial as M_1 outperforming M_2 ($\overline{Z_d} > 0$) on a particular catchment, whereas a failure is M_2 outperforming M_1 ($\overline{Z_d} < 0$). Let V be a random variable defined as the number of catchments where M_1 outperforms M_2 (successes). Thus the probability of v instances where M_1 outperforms M_2 is a binomial distribution $B(c, p)$ and given as;

$$P\{V = v\} = \binom{c}{v} . p^v . (1 - p)^{(c-v)} \tag{8}$$

A hypothesis test can be formed for a predefined significance level e.g. $\alpha = 5\%$. The observed number of successes v, is compared to the critical interval as defined as $v \leq B(c, p)_\alpha$ where subscripts α signify the quantile of the binomial distribution. A p-value can be derived

from Eq 8 in order to determine the probability of a specified v falling within critical interval α. If the observed v falls within the critical interval, then the null hypothesis can be rejected such that there is a difference in model performance. If v does not fall within the critical interval then the null hypothesis can be accepted such that there is no difference in model performance. Figure 3 is an example of the method for a one-tailed example let $c = 27$ and $p = 0.5$, then to achieve a significant difference in model performance at $\alpha = 5\%$ significance level, M_1 would have to outperform M_2 in 18 out of 27 catchments as indicated by the white section of Figure 3.

<Figure 3 >

CASE STUDY: THE THAMES CATCHMENT

The two hydrological model comparison tools developed in the previous sections were tested using two conceptual rainfall-runoff models on a set of $c = 27$ gauged catchments located within the Thames catchment.

Model description

The two models used for this case study are URMOD (M_2) (Fidal 2019) and DAYMOD (M_1) (Kjeldsen et al. 2005). URMOD is an extension of DAYMOD containing an urban runoff framework to account for urban land-use, resulting in a nested model structure where M_1 is a simpler version of M_2. Both models are a lumped-conceptual parameter-parsimonious rainfall-runoff models with URMOD having eight calibrated parameters whereas DAYMOD has seven. The models represent two main processes (i) infiltration and runoff and (ii) channel routing.

The infiltration and runoff generation is based on a conceptual soil column approach, such that the precipitation that does not infiltrate is turned into direct runoff. The runoff generation in DAYMOD is dependent on the soil moisture in the conceptual soil column such that as the column fills, more runoff is generated. The runoff generation in URMOD is split into two contributions, one from the rural areas and the second from the urban areas. Runoff

generation from the rural areas is the same as DAYMOD, whereas urban runoff generation is determined via a calibrated parameter.

The second process within the models is the channel routing which is based on parallel linear reservoir. Routing from the rural areas is achieved via a linear reservoir with a proportion of the runoff routed through a local baseflow reservoir before being routed through a surface flow reservoir. In contrast, the proportion of runoff designated as surface flow is just routed through the surface flow reservoir. The baseflow and surface flow is then combined to be the rural runoff at the catchment outlet. The urban routing within URMOD routes the urban runoff through a separate surface flow reservoir, and then combines with the rural runoff to become total runoff at the catchment outlet.

The two models require observed rainfall, runoff and potential evaporation, in order to calibrate the eight parameters (DAYMOD has seven calibrated parameters). Each model is calibrated by first selecting initial conditions and parameters, followed by calibration of the optimal parameters using the shuffled evolution complex algorithm (Duan et al. 1993).

Catchment Selection

An initial set of 112 catchments were assembled from within the Thames catchment for which long-term daily rainfall and runoff data are available from the National River Flow Archive (NRFA). This initial set was reduced to a subset of 27 catchments based on the condition that the fraction of urban land cover had to be larger than 5% of the catchment to ensure a meaningful comparison of URMOD and DAYMOD. Furthermore, each catchment needed good quality data for a 30-year period 01/01/1980 to 31/12/2009. The resulting 27 catchments ranged in size from 21.8 km^2 to 904 km^2 with fractional urban land cover values ranging from 5.34% to 54.75%. In Figure 4 the 27 catchments are highlighted in grey.

<Figure 4 >

The hydro-meteorological data used in this study consist of: catchment average daily precipitation (i), average daily river flow (q_{obs}), and daily potential evaporation data (E_p). Runoff data at a daily time step were acquired from the National River Flow Archive (NRFA)

spanning 30-years from 01/01/1980 up to 31/12/2009. The precipitation data were obtained from the CEH-GEAR data set (Keller et al. 2015) covering the same 30-year period. Finally, evaporation data were obtained from the Climate, Hydrology and Ecology research support system (CHESS) (Robinson et al. 2016). The runoff data were quality-controlled by removing the missing data rather than estimating values, checks were made to ensure that there were no major gaps of multiple weeks worth of data missing.

One important criteria for the urban model is determining the percentage of urban land-use in a catchment. For this study the URBEXT_{2000} catchment descriptor (Bayliss et al. 2006) was used, where the subscript 2000 denotes that the $50m$ x $50m$ land-cover data used to construct the index refers to land-use data from the period between the years of 1998-2000. URBEXT_{2000} uses a contribution of both urban and sub-urban land-cover classes, with the urban land-cover consisting of roofs, roads and man-made structures, whereas the sub-urban section is a mix of vegetation and semi-built up areas, only half of the sub-urban section is contributed to URBEXT_{2000} as it is assumed that half of the sub-urban section is made up of vegetation such as gardens or parks (Bayliss et al. 2006).

Selection of performance criteria

In order to apply the two methods, performance criteria needs to be selected. Two performance criteria are selected for this study the Nash-Sutcliffe efficiency statistic (NSE) (Nash and Sutcliffe 1970) and the volumetric efficiency (VE). Consider a time series of observed runoff $q_{obs}(t), t = 1, ..., n$, and the accompanying simulated runoff $q_{sim}(t), t = 1, ..., n$ obtained from a calibrated rainfall-runoff model. The Nash-Sutcliffe efficiency statistic (NSE) statistic is defined as

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{n}(q_{obs}(t) - q_{sim}(t))^2}{\sum_{t=1}^{n}(q_{obs}(t) - \bar{q}_{obs}(t))^2} \tag{9}$$

The range of possible NSE values spans from $-\infty$ to one, with a value of one indicating perfect fit, i.e $q_{sim} = q_{obs}$ for all n observations. This criteria was selected because of its

widespread use in hydrology. An often cited problem with this criteria is that it is sensitive to extremal events due to squaring differences as discussed by Krause et al. (2005) and Legates and McCabe (1999). The second criteria to be used is the volumetric efficiency (VE) (Criss and Winston 2008) defined as

$$VE = 1 - \frac{\sum_{t=1}^{n} |q_{sim}(t) - q_{obs}(t)|}{\sum_{t=1}^{n} q_{obs}(t)} \tag{10}$$

The range of possible VE values spans from 0 to 1, with a value of 1 indicating perfect fit. This criteria was chosen as it attempts to overcome the shortcomings of the NSE. One advantage is that it is based on absolute values as opposed to squaring, which the NSE does, and so applies an equal weight to each flow comparison (Criss and Winston 2008). Whilst only two performance criteria are used in this study it is possible to use more when comparing models.

Experiment setup

The two methods presented in the Model comparison techniques section were used in an attempt to compare the performance of the two rainfall-runoff models URMOD (M_2) and DAYMOD (M_1) when applied to 30-years of observed hydrological data on c =27 catchments in the Thames basin. This case study is interested in exploring the potential benefits of applying an urban framework to a rural model in order to account for urbanisation in catchments. Two comparisons will be undertaken. The first when both models are calibrated on each of the 27 catchments in turn with 1-year of calibration data, which will result in $n = 30$ performance criteria (for each criteria) and parameter sets for each catchment. The second comparison will involve both models being calibrated on the 27 catchments with 2-years of calibration data, which will result in $n = 15$ performance criteria (for each criteria) and parameter sets for each catchment. These calibration periods were chosen in order to obtain suitably large number of performance criteria for the jackknife variance.

The paired t-test method was applied to each of the 27 catchments in order to compare the

performance of the two models on each catchment in turn. A two-tailed test is chosen because it is important to determine if there is a significant difference in performance between the models, as opposed to a singular model outperforming the other. However for the binomial hypothesis test a one-tailed test will be used, with alternative hypothesis that URMOD outperforms DAYMOD. A one-tailed test is chosen in this case to investigate if there is a significant difference between the two models in favor of URMOD across a wide range of urban catchments.

RESULTS

Assessing performance of individual catchments via paired t-test

This section will explore the difference in performance of the two models at the individual catchment level, using the paired t-test. The results will be presented such that a positive difference in performance criteria ($\overline{Z_d} > 0$) indicates model M_2 (URMOD) performed better than model M_1 (DAYMOD). In contrast negative difference ($\overline{Z_d} < 0$) indicates that model M_1 (DAYMOD) performed better than model M_2. The 95% confidence intervals of $\overline{Z_d}$ are calculated for each catchment using Eq 3 and then subsequently plotted. If the confidence intervals cross zero, then the null hypothesis of equal performance between M_1 and M_2 can be accepted, whereas if the interval does not cross zero the null hypothesis is rejected such that there is a significant difference between the models.

1-year calibration period results

Figure 5 shows the results of difference in performance of M_2 and M_1 when both models are calibrated on 1-year, and the performance assessed on the 29-year validation periods, thus $n = 30$. The left hand figure show the difference in performance criteria when using the NSE performance criteria, and the right hand side is the difference in performance criteria when using the VE performance criteria. The circles indicate M_2 has a larger performance criteria than M_1 i.e $\overline{Z_d} > 0$, whilst the triangles show the reverse, i.e $\overline{Z_d} < 0$. The lines indicate the 95% confidence interval such that if they cross zero the null hypothesis of equal

performance can be accepted.

<Figure 5 >

When the NSE is applied M_2 outperformed M_1 on 14 catchments (out of 27), and when the VE is applied M_2 outperformed M_1 on 19 catchments. All of the confidence intervals for both NSE and VE have crossed zero indicating that the null hypothesis of equal performance cannot be rejected.

When the difference in performance of the models is explored with respect to area and percentage of urbanisation no trend was seen. However when exploring the difference in soil type of the rural section of the catchment (defined using BFIHOST) it shows that M_1 performs better on catchments with more permeable soil, whilst the reverse is true for M_2 when the NSE results are used. However when the VE results are used this effect is less apparent.

2-year calibration period results

Figure 6 shows the results of difference in performance of M_2 and M_1 calibrated on 2-years, with validation period of 28-years, thus $n = 15$. Similar to the results presented in Figure 5, the circles indicate that $\overline{Z_d} > 0$, whilst the triangles represent $\overline{Z_d} < 0$. The left hand figure is the difference in performance when the NSE is applied, while the right hand side is the difference in performance criteria when the VE is applied. Again the lines indicate the 95% confidence intervals.

<Figure 6 >

When the NSE criteria is applied M_2 outperformed M_1 in 15 out of 27 catchments, whereas when the VE is applied M_2 outperformed M_1 on 14 out of 27 catchments. Again, all confidence intervals crossed zero indicating the null hypothesis of equal performance cannot be rejected. Similar to the results obtained using 1-year calibration, no trend is appeared when considering performance across urbanisation. Again M_1 performed better on catchments with permeable soils, with M_2 performing better on catchments with less permeable soils.

435 **Assessing performance of collective catchments via Binomial hypothesis method**

436 This section will explore the difference in performance of the two models collectively
437 across all 27 catchments, using the binomial distribution test. The hypothesis test is a one-
438 tailed test such that the alternative hypothesis is defined as H_1: M_2 performs statistically
439 significantly better than M_1 ($p > 0.5$). A $\alpha = 5\%$ significance level is chosen, such that if
440 the p-value obtained is less than 0.05, the null hypothesis H_0 can be rejected whilst a larger
441 p-value indicates the null hypothesis H_0 (M_1 performing the same or better than M_2) cannot
442 be rejected. Four different hypothesis tests are formulated each of them with the same null
443 and alternative hypothesis. The tests differ by the calibration year (1-year or 2-year) and
444 the performance criteria used (NSE or VE). The results of all four binomial hypothesis tests
445 are shown in Table 1.

446 <Table 1 >

447 Table 1 shows that the null hypothesis can be accepted for three cases, NSE based on
448 1-year,2-year and VE for 2-year. However the null hypothesis can not be rejected for the
449 1-year calibration with VE as the performance criteria. The acceptance of H_0 means that
450 M_2 (the urban model) performed significantly better than M_1 across the 27 catchments.

451 ## DISCUSSION

452 The results and methodology presented in this study raises a number of issues that need
453 further discussion. Two different performance criteria, the NSE and VE were selected for
454 this study and both led to two different conclusions once the binomial hypothesis test was
455 applied. This further highlights Legates and McCabe (1999) conclusion that the choice of
456 performance criteria needs to be made clear prior to application. In this study two lumped
457 conceptual models (DAYMOD and URMOD) were chosen, with URMOD accounting for
458 urban surfaces and DAYMOD being a more simple nested version of the URMOD model
459 structure. But the flexibility of the hypothesis tests and calibration methodology posted
460 within this paper can be applied to any comparison of models in need of calibration and
461 validation.

As shown in the results section, using the paired t-test approach the difference in model performance was not statistically significant on any of the 28 test catchments (as indicated by the significance lines not crossing zero). This could be because the performance of the two models being indistinguishable, or could indicate that the 28 or 29-years of lumped data still creates too much variation in performance criteria to distinguish model performance. Hence further research is needed to determine if this method is a viable hydrological comparison tool. In the application of the jackknife calibration and validation method a common record length was assumed (30-years), with a common calibration sub-period selected (1-year and 2-year). However, a varying record-length for each catchment can be used. Varying sub-period length can also be used for different catchments but the calibration sub-periods length has to be consistent for a singular catchment. The method can be used to split the record length into variable hydrological periods, similar to the differential split-sample test described by Klemeš (1986), thus creating multiple differential split-samples. Hence a comparative analysis can then be conducted between the different hydrological periods. The binomial hypothesis test can also be applied in order to explore the performance between these different hydrological periods. By splitting the periods into two sub-periods, a binomial hypothesis test can be applied to both.

One clear advantage of the proposed binomial test is the ability to assess if a model is significantly better performing across a large number of catchments. This is a simple-to-use methodology that applies commonly used performance criteria. The binomial approach is also flexible with the jackknife calibration and validation method, rather than comparing multiple catchments the binomial method can be applied to a singular catchment, such that the trials would be denoted as individual years or certain events.

Whilst conflicting conclusions between the methods may seem like an issue, the purpose of both tests are to answer different questions. The paired t-test showed no significance between model performance at the individual catchment level, but the binomial hypothesis test showed that one model performed statistically better across a number of catchments in

one out of the four hypothesis tests. One reason for this is that the performance of the urban model is indistinguishable from the rural model when comparing models on the individual catchment level. However performance is significantly better when comparing performance across a group of catchments but it is still difficult to characterise with such similar models.

DATA AVAILABILITY STATEMENT

The following data used during the study were provided by a third-party.

- Rainfall data (Keller et al. 2015).
- Evaporation data (Robinson et al. 2016).
- River flow data (NRFA 2018).

Direct requests for these materials may be made to the provider as indicated in the "Acknowledgments."

The following models and code generated used during the study are available from the corresponding author by request.

- URMOD (hydrological model).
- Jackknife calibration/validation code.

CONCLUSION

This paper presented two new easy-to-use techniques for comparing the performance of rainfall-runoff models, as well as presenting a more robust methodology to calibrate and validate rainfall-runoff models. The paired t-test is used to determine comparative model performance for a single catchment, whilst the binomial hypothesis test considers model performance across a group of catchments. The results showed that when comparing URMOD and DAYMOD no significant differences were obtained on a catchment by catchment level when the t-test was applied. This shows that simply introducing an urban surface to account for urbanisation was not enough to have a significant effect at the individual catchment level. When the Binomial hypothesis test was applied it showed that when the NSE was applied

there was either no difference or DAYMOD performed better across a wide range of catchments. However when the VE was applied for a 1-year calibration URMOD did perform statistically significantly better than DAYMOD.

The purpose of this paper was to show that applying simple statistical methods can add interpretive power when comparing model performance. Poor performing models can appear to perform well when performing a simple split-sample test and applying performance criteria, reflecting the subjective conclusions that can be drawn from simply reporting a single performance criteria. The new tools developed within this paper allow a more rigorous analyse based on commonly accepted statistical hypothesis tests and so have the potential to improve model performance analyse. However whilst these techniques do add a more robust method to test model performance, it is recommenced that these techniques should be used alongside other performance methods such as graphical analyses (hydrographs) where possible to ensure maximum robustness of models.

Acknowledgments

The authors would like to thank the three anonymous reviewers for helpful comments which helped improve the paper. The National River Flow Archive (NRFA) for providing access to hydrological data and catchment shapefiles. Funding from the Engineering and Physical Sciences Research Council (EPSEC) (grant 1552086) is acknowledged. The NERC funded POLLCURB project for providing access to the hydrological and land-use data used in this study (NE/K002317/1).

REFERENCES

Addor, N. and Melsen, L. (2019). "Legacy, rather than adequacy, drives the selection of hydrological models." *Water Resources Research*, 55(1), 378–390.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M., and Valéry, A. (2009). "Crash tests for a standardized evaluation of hydrological models." *Hydrology and Earth System Sciences*, 10(13), 1757–1764.

Anh, N. L., Boxall, J., Saul, A., and Willems, P. (2010). "An evaluation of three lumped conceptual rainfall-runoff models at catchment scale." *Proceedings of the 3rd International Symposium on British Hydrological Society, Newcastle, UK, July*.

Bayliss, A., Black, K., Fava-Verde, A., and Kjeldsen, T. (2006). "URBEXT2000 - a new FEH catchment descriptor. calculation, dissemination and application." *Report No. R&D FD1919/TR*, Department for Environment Food and Rural Affairs, CEH wallingford.

Beven, K. J. (2011). *Rainfall-runoff modelling: the primer*. John Wiley & Sons, Chichester.

Bouffard, J.-S. (2014). "A comparison of conceptual rainfall-runoff modelling structures and approaches for hydrologic prediction in ungauged peatland basins of the james bay lowlands." Ph.D. thesis, Carleton University Ottawa, Carleton University Ottawa. Retrieved from https://curve.carleton.ca/7ed3b15f-bfff-4027-9a87-f314602d7a1a.

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F. (2012). "Crash testing hydrological models in contrasted climate conditions: An experiment on 216 australian catchments." *Water Resources Research*, 48(5).

Criss, R. E. and Winston, W. E. (2008). "Do nash values have value? discussion and alternate proposals." *Hydrological Processes*, 22(14), 2723–2725.

Donnelly-Makowecki, L. and Moore, R. (1999). "Hierarchical testing of three rainfall–runoff models in small forested catchments." *Journal of Hydrology*, 219(3), 136–152.

Duan, Q., Gupta, V. K., and Sorooshian, S. (1993). "Shuffled complex evolution approach for effective and efficient global minimization." *Journal of Optimization Theory and Applications*, 76(3), 501–521.

561 Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, Vol. 38. Society
562 for Industrial and Applied Mathematics.

563 Ewen, J. (2011). "Hydrograph matching method for measuring model performance." *Journal*
564 *of Hydrology*, 408(1), 178–187.

565 Ewen, J. and O'Donnell, G. (2012). "Prediction intervals for rainfall–runoff models: raw
566 error method and split-sample validation." *Hydrology Research*, 43(5), 637–648.

567 Fidal, J. (2019). "Investigating the impact of urbanisation on rainfall-runoff models." Ph.D.
568 thesis, University of Bath, University of Bath.

569 Fleming, S. (2009). "An informal survey of watershed model users: preferences, applications,
570 and rationales." *Streamline Watershed ManageBull*, 13(1), 32–35.

571 Gharari, S., Hrachowitz, M., Fenicia, F., and Savenije, H. (2013). "An approach to identify
572 time consistent model parameters: sub-period calibration." *Hydrology and Earth System*
573 *Sciences, 17, 2013*.

574 Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). "Decomposition of the
575 mean squared error and nse performance criteria: Implications for improving hydrological
576 modelling." *Journal of Hydrology*, 377(1-2), 80–91.

577 Jones, D. A. and Kay, A. L. (2007). "Uncertainty analysis for estimating flood frequencies for
578 ungauged catchments using rainfall-runoff models." *Advances in Water Resources*, 30(5),
579 1190–1204.

580 Kavetski, D., Kuczera, G., and Franks, S. W. (2006). "Bayesian analysis of input uncertainty
581 in hydrological modeling: 1. theory." *Water Resources Research*, 42(3).

582 Keller, V., Tanguy, M., Prosdocimi, I., Terry, J., Hitt, O., Cole, S., Fry, M., Morris, D.,
583 and Dixon, H. (2015). "CEH-GEAR: 1 km resolution daily and monthly areal rainfall
584 estimates for the UK for hydrological and other applications." *Earth System Science Data*,
585 <https://doi.org/10.5285/5dc179dc-f692-49ba-9326-a6893a503f6e>.

586 Kirchner, J. W., Hooper, R. P., Kendall, C., Neal, C., and Leavesley, G. (1996). "Testing
587 and validating environmental models." *Science of the Total Environment*, 183(1-2), 33–47.

588 Kjeldsen, T., Stewart, E., Packman, J., Folwell, S., and Bayliss, A. (2005). "Revitalisa-
589 tion of the FSR/FEH rainfall-runoff method." *Report no.*, Defra R&D Technical Report
590 FD1913/TR, CEH Wallingford.
591 Klemeš, V. (1986). "Operational testing of hydrological simulation models." *Hydrological*
592 *Sciences Journal*, 31(1), 13–24.
593 Kohavi, R. et al. (1995). "A study of cross-validation and bootstrap for accuracy estimation
594 and model selection." *IjcaI'95 Proceedings of the 14th international joint conference on*
595 *Artificial intelligence*, Vol. 2, Montreal, Canada, 1137–1145.
596 Krause, P., Boyle, D., and Bäse, F. (2005). "Comparison of different efficiency criteria for
597 hydrological model assessment." *Advances in Geosciences*, 5, 89–97.
598 Legates, D. R. and McCabe, G. J. (1999). "Evaluating the use of goodness-of-fit measures in
599 hydrologic and hydroclimatic model validation." *Water Resources Research*, 35(1), 233–
600 241.
601 Mishra, S. (2009). "Uncertainty and sensitivity analysis techniques for hydrologic modeling."
602 *Journal of Hydroinformatics*, 11(3-4), 282–296.
603 Nash, J. E. and Sutcliffe, J. V. (1970). "River flow forecasting through conceptual models
604 part I-A discussion of principles." *Journal of Hydrology*, 10(3), 282–290.
605 NRFA (2018). "National river flow archive, NERC CEH, Wallingford,
606 <https://nrfa.ceh.ac.uk/>.
607 Pappenberger, F. and Beven, K. J. (2006). "Ignorance is bliss: Or seven reasons not to use
608 uncertainty analysis." *Water Resources Research*, 42(W05302).
609 Pechlivanidis, I., Jackson, B., and McMillan, H. (2010). "The use of entropy as a model diag-
610 nostic in rainfall-runoff modelling." *iEMSs 2010. International Congress on Environmental*
611 *Modelling and Software. 5 July, Ottawa, Canada*, Vol. 2.
612 Quenouille, M. H. (1956). "Notes on bias in estimation." *Biometrika*, 43(3/4), 353–360.
613 Refsgaard, J. C. (1997). "Parameterisation, calibration and validation of distributed hydro-
614 logical models." *Journal of Hydrology*, 198(1), 69–97.

615 Refsgaard, J. C. and Knudsen, J. (1996). "Operational validation and intercomparison of
616 different types of hydrological models." *Water Resources Research*, 32(7), 2189–2202.
617 Robinson, E., Blyth, E., Clark, D., Comyn-Platt, E., Finch, J., and Rudd, A. (2016).
618 "Climate hydrology and ecology research support system meteorology dataset for
619 great britain (1961-2015)[chess-met], <https://doi.org/10.5285/8baf805d-39ce-4dac-b224-
620 c926ada353b7>.
621 Santos, C., Almeida, C., Ramos, T., Rocha, F., Oliveira, R., and Neves, R. (2018). "Using
622 a hierarchical approach to calibrate swat and predict the semi-arid hydrologic regime of
623 northeastern brazil." *Water*, 10(9), 1137.
624 Schaefli, B. and Gupta, H. V. (2007). "Do Nash values have value?." *Hydrological Processes*,
625 21(15), 2075–2080.
626 Seibert, J. (2003). "Reliability of model predictions outside calibration conditionspaper pre-
627 sented at the nordic hydrological conference (røros, norway 4-7 august 2002)." *Hydrology
628 Research*, 34(5), 477–492.
629 Seibert, J., Vis, M. J., Lewis, E., and van Meerveld, H. (2018). "Upper and lower benchmarks
630 in hydrological modelling." *Hydrological Processes*, 32(8), 1120–1125.
631 Selle, B. and Hannah, M. (2010). "A bootstrap approach to assess parameter uncertainty in
632 simple catchment models." *Environmental Modelling & Software*, 25(8), 919–926.
633 Shen, Z., Chen, L., and Chen, T. (2012). "Analysis of parameter uncertainty in hydrological
634 and sediment modeling using glue method: a case study of swat model applied to three
635 gorges reservoir region, china." *Hydrology and Earth System Sciences*, 16(1), 121–132.
636 Vogel, R. M. and Sankarasubramanian, A. (2003). "Validation of a watershed model without
637 calibration." *Water Resources Research*, 39(10).
638 Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S. (2003). "A shuffled complex
639 evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic
640 model parameters." *Water resources research*, 39(8).
641 Weglarczyk, S. (1998). "The interdependence and applicability of some statistical quality

642 measures for hydrological models." *Journal of Hydrology*, 206(1-2), 98–103.

643 Xu, C. (1999). "Operational testing of a water balance model for predicting climate change

644 impacts." *Agricultural and Forest Meteorology*, 98, 295–304.